AN EFFICIENT ALGORITHM FOR MULTIPLE POLONY DETECTION

H. Leandro Cortés[*]

Department of Computer Science
The University of Chicago,
Chicago, IL 60637

Gregory Snyder[†]

Department of Human Genetics
The University of Chicago
Chicago, IL 60637

ABSTRACT

We present an algorithm for automatic detection of solid-phase (polony) Polymerase Chain Reaction (PCR) objects. The goal is to detect the location and size of each polony present in an image. Using a statistical model for an image of a polony, we are able to weigh different hypothesis, including arrangements of multiple overlapping objects. The algorithm uses a coarse-to-fine approach. A coarse version of the model is used to find candidate object locations and sizes with a low computational cost. More accurate versions of the full model, are then used to test for object presence only at the candidate locations.

Index Terms— Object detection, image analysis, data models, biomedical image processing.

1. INTRODUCTION

We present an algorithm for multiple object detection and test it on images of an application of solid-phase (polony) PCR to gene expression analysis [1]. The ability of polony PCR to detect single molecules of DNA makes it suitable for studying gene expression in single cells. Gene expression is the process by which genes in chromosomes, the information-carrying molecules of the cell, are converted into proteins, the functional molecules of the cell. Not every gene is expressed in every cell, and one sense of the identity of a cell can be defined as the set of genes it expresses. It is increasingly being appreciated that cell-to-cell variation in gene expression ("noise") plays a role in processes such as development and aging and may be involved in diseases such as cancer [2, 3, 4, 5].

Solid-phase PCR is performed in a polyacrylamide gel attached to a glass microscope slide. One PCR primer is crosslinked to the gel and the other is free to diffuse. PCR-amplified sequences are detected by hybridization of fluorescently-labeled complementary oligonucleotide probes. Labeled slides are imaged on a micro-array scanner which raster-scans a focused laser beam over the slide and detects fluorescence with a photomultiplier tube. This method enables detection of single DNA molecules by making many copies of them and restricting the copies spatially to the vicinity of the starting molecule, separated from other molecules

A few methods are known by the authors to be used for polony detection [6, 7]. Usually morphology toolboxes from packages as Matlab and lab-made image processing software is used for this task. Model based methods, pretrained classifiers and some image pre-processing techniques have been used in similar images as those of single fluorescent particles[8, 9, 10]. However, most of these methods are unable to describe groups of pixels and lack the flexibility to test different explanations for the data when multiple objects are present in a scene.

The main contribution of this paper is a model-based method for multiple polony detection. The algorithm relies on a statistical model of the data, providing the ability to formulate well defined hypothesis tests for the presence of objects and the flexibility to deal with multiple polonies located in a very small region. Other advantages include little parameter tuning and efficient computation. A more complex version of the method presented here, applied to a different setting including motion is presented in [11]. Similar ideas of multiple object detection in the analysis of static images can be found in [12, 13].

2. STATISTICAL MODEL

The image of a polony is modeled by a bell-shaped function of variable width, location, and amplitude, to which are added a constant background and noise. The intensity at location $x = (x_1, x_2)$ is modeled as $I(x) = G_\omega(x) + C + \zeta_{\sigma_{noise}}$ where $G_\omega(x)$ is a bell-shaped function with parameter ω, C is a baseline and $\zeta_{\sigma_{noise}}$ is i.i.d noise that follows a normal distribution with mean zero and standard deviation σ_{noise}.

Multiple objects might be present in the same region producing a polony configuration. The general model of the image intensity $I(x)$ is then given by

$$I(x) = C(x) + \sum_i G_{\omega_i}(x) + \zeta_{\sigma_{noise}} \quad (1)$$

[*]Supported in part by Burroughs Wellcome Fund Interfaces 1001774 and NSF ITR DMS-0219016

[†]Supported by NIH Ruth Kirchstein NRSA fellowship F32 GM075503

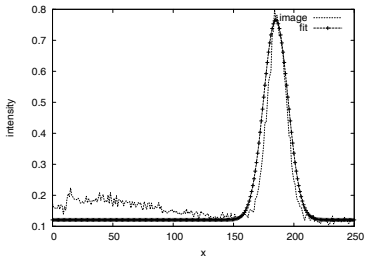


Fig. 1. A profile of the intensity values of a polony and the fitted model.

where i is an index over polonies and $C(x)$ is the background intensity at x

We assume that $C \geq 0$ is smoothly changing over the image, so that when a small region is considered, it is reasonable to assume that C is constant. Thus, for each object ω, we may have a different constant value C_ω for the immediate region around μ. We define

$$G_\omega(x) = Ae^{-\|x-\mu\|^2/(2r^2)} \tag{2}$$

where $\omega = (\mu, r, A, C)$ is the state vector composed of the location μ, the width r and the maximum intensity A.

If we assume that the intensity values of a polony are negligible at distance greater than δr_i from μ_i, we can define a simpler model for the image intensities $I(x)$:

$$I(x) = C + \sum_{i:|\mu_i - x| \leq \delta r_i} G_{\omega_i}(x) + \zeta_{\sigma_{noise}} \tag{3}$$

where C is assumed to be the same constant for the nearby polonies.

Variations of the Gaussian model (2) have been used in single fluorescent particle applications [8, 10]. In these cases, the Gaussian model is used as an approximation to the point spread function of the microscope. There is no strong physical justification for the model presented here, other than shape resemblance (see Figure-1). However, this model is enough for the purpose of this application, that is, a description of the polonies in terms of location, relative size and intensity.

3. DETECTION PROCEDURE

Given an image I, detection consists of finding a set of k objects $(\omega_1, \omega_2, ..., \omega_k)$, for an unknown integer $k \leq K$, which maximizes the posterior

$$\arg\max P\left((\omega_i)_{i=1}^k \mid I\right), \tag{4}$$

Solving (4) is a difficult optimization problem over an unknown number of variables. To approximate the solution we use a coarse-to-fine approach. First, we define a simplified binary oriented edge based model, which allows us to ignore the amplitude state variable A and the baseline C. This leads to a simple and efficient edge-based test that eliminates the vast majority of possible object locations and bounds the number of objects present in the image.

Subsequently, the precise shape parameters of candidate polonies are estimated using the full static data model (3) which is adapted to the number of candidate detections in a region. This step can also eliminate detections if the estimated amplitude A for any of the candidates is below a threshold. All the tests are standard statistical tests derived from the hypothesized data distribution using conservative thresholds to avoid losing true positives.

3.1. Edge based models

In the first step of detection we aim at deriving a test that is invariant to amplitude and baseline variations, in order to avoid the computationally intensive task of actually estimating these variables, for every candidate location. Detection thus starts using information only from locations where there is large variation in image intensity. We focus our attention on the spatial gradient of the polony template (1). Therefore, high values will occur at the circles around the centers of the objects at radius given by r_i.

Thus, the first coarse model for polonies is defined as a set of templates of circles of radius $r \in \Sigma$, where Σ is a finite set of feasible polony sizes. The templates are defined in terms of binary features labeled according to orientation. Accordingly, the input image is transformed into a matrix of binary features. At each pixel, the presence and label of a feature is determined by thresholding the magnitude of the gradient of the image. Using low thresholds, a large number of features are likely to occur on circles, indicating the presence of a candidate object.

The first step of detection reduces to circle detection in an array of binary oriented edges, which can be solved efficiently using the Hough transform [14]. A simple probability computation sets a conservative threshold on the feature counts to avoid false negatives, yielding an initial list of candidate locations and sizes for the vesicles.

To avoid unnecessary processing we use a mask to define a region of interest (ROI) for object detection. All candidate detections that are outside the ROI are automatically discarded.

Assuming constant size among different images, the ROI is implemented as a binary mask that covers the ellipsoid that contains the gel (see Figure-3a). The mask is aligned with the input image I, by finding the displacement $d \in D$ that minimizes the sum $\sum_{x \in E} D_M(x+d)$, where E is the set of pixels selected after thresholding the magnitude of the gradient of I and $D_M(x)$ is the distance from each pixel to the nearest pixel on the edge of the ROI. Thus, D_M is the distance transform of the edges of the mask [15], which can be pre-computed offline because it does not depend on I.

3.2. Refined models

At the candidate locations that survived the previous test we compute the optimal values of the state variables using the full gray level model which may involve several polonies simultaneously. Assume first that in a region of interest R of size δr around a given detection (μ, r) there are no other detections. We try to find the optimal value for the state variables in the image model defined in (3),

$$\arg\min_{\mu,r,A,C} \sum_{x\in R} [I(x) - C - Ag_{\mu,r}(x)]^2. \quad (5)$$

We optimize the likelihood in (5) by iterating between pairs of variables (μ, r), and (A, C). For fixed values of μ and r, we can estimate the global optimum of A and C using linear regression, with $g_{\mu,r}(x), x \in R$ acting as the independent variable. Then, for fixed A, C, the cost is optimized in (μ, r) using a conjugate gradient algorithm[16]. The procedure is iterated until the relative error is close to 1 or the number of iterations is too high. A standard t-test [17] at some significance level α is then used to determine whether the amplitude A is zero, in which case we reject the detection ω.

3.3. Object configurations

When the region R includes $K > 1$ polonies, the full model given in equation (3) is used. In this case we have to minimize:

$$J(\omega_1, \ldots, \omega_K) = \sum_{x\in R} [I(x) - C - \sum_{k=1}^{K} A_k g_{(\mu_k, r_k)}(x)]^2 \quad (6)$$

where k runs over detections ω_k such that $\mu_k \in R$

The methods used before apply in the multiple polony framework. Now for fixed $\mu_k, r_k, k = 1, \ldots, K$ perform a multiple linear regression to solve for $C, A_k, k = 1, \ldots, K$ and the conjugate gradient method is implemented simultaneously on all the remaining variables. The significance of each fit is determined using a standard t-test to determine whether A_k is zero.

In practice, region R is preferred to be small, containing a few, close objects. The computation time, specially for the conjugate gradient algorithm, increases considerably fast, as a function of the area of R. Thus, regions are bound to have a maximum size s_{MAX}. Let s_R to be the radius of the region R and L_R the list of objects ω inside R. Starting with object ω_1, we set $L_R \leftarrow \{\omega_1\}$ and $s \leftarrow 4r_1$. For each object ω' with center inside R, we set $L_R \leftarrow L_R \cup \omega'$ and increase s to include the new object ($s \leftarrow \min\{s + 2r', s_{MAX}\}$). The (multiple) Gaussian fitting procedure is then applied using all objects inside R, but the parameter update is accepted only for those candidate objects that are not rejected and fall completely inside R. Thus, the process must be repeated with new regions, for each of the candidate objects that did not fall completely inside R.

4. RESULTS

The algorithm was implemented in C++. It is being used with images of an application of solid-phase (polony) PCR to gene expression analysis. Gray level images of size 1726×2485 are recorded as described in section-1 and stored in TIFF files using a two-byte-per-pixel format. One image usually contains several hundreds of polonies (see Figure-3a,b). Using one processor of an Intel Core 2 2GHz, image processing takes $0.8 \pm .01$s, the coarse detection step takes 0.006 ± 0.004s per object and the refined detection takes 0.075 ± 0.03s per object.

The gradient estimation is done with an implementation of the Deriche filter [18]. The size parameter α of the Deriche filter was set to produce appropriate edges in the expected scale of the polonies. Low thresholds for the magnitude of the gradient and the vote from the Hough transform were set using a few sample images.

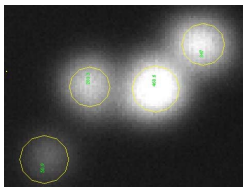


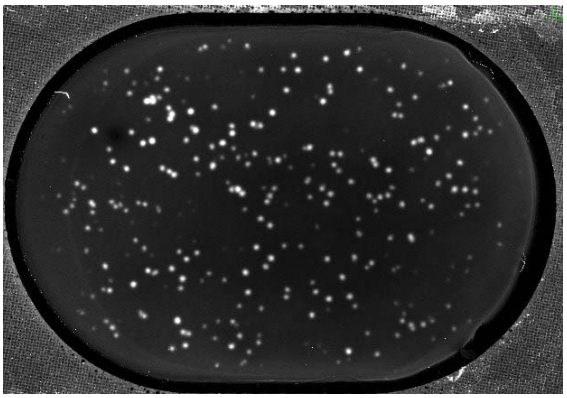
Fig. 2. Detail of a configuration of polonies

An example of the output produced by the program is shown in Figure-3. The images (a-b) show a slide with a few hundreds of detections made by the program. The zoomed image (Figure-2) shows a configuration of multiple polonies.

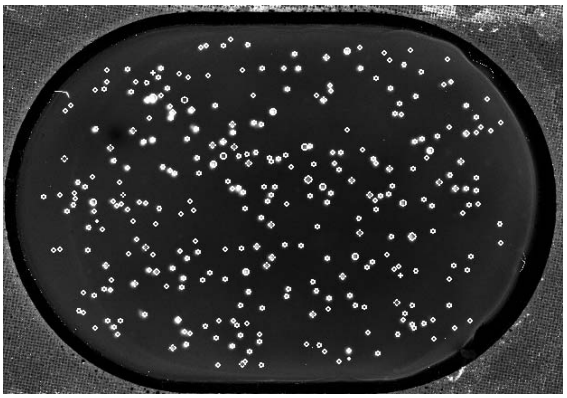
To measure performance of the algorithm, we did a visual inspection of five images and found 1244 objects that looked like polonies. The program was able to detect 1150 polonies, produced 13 false positives and missed 94 polonies ($\ll 7.5\%$), distributed as 33 overlapped polonies and 61 very dim polony-like objects. The missed polonies that were partially occluded are known to be related to a problem in the coarse detection step, that assumes a single object, being unable to handle two objects with overlapping of more than half of one object. False positives were produced by dirt on the gel. Dirt was usually small and very bright in all channels when using multi color registration. Thus, most of these false detections could be eliminated with help of size and/or amplitude information from detections in multiple channels.

5. ACKNOWLEDGEMENTS

The authors thank Professor Yali Amit (Dept. Statistics and Dept. of Computer Science at the University of Chicago) for advice, thorough discussions and feedback on the draft. Dr. Eric Schwartz (Dept. of Neurobiology, Pharmacology and Physiology) for discussions and testing of the core of the software, developed during a preceding work with video data.



(a) original



(b) detections

Fig. 3. Polony detection example. (a) original image: the dark oval is the boundary of the gel, the small dots are the polonies (b) image + detections.

6. REFERENCES

[1] R. D. Mitra and G. M. Church, "In situ localized amplification and contact replication of many individual dna molecules," *Nucleic Acids Research*, vol. 27, no. 24, pp. e34, 1999.

[2] M. S. Samoilov, G. Price, and A. P. Arkin, "From fluctuations to phenotypes: the physiology of noise," *Science STKE*, 2006.

[3] R. Bahar, C. H. Hartmann, K. A. Rodriguez, A. D. Denny, R. A. Busuttil, M. E. T. Doll, R. B. Calder, G. B. Chisholm, B. H. Pollock, C. A. Klein, and J. Vijg, "Increased cell-to-cell variation in gene expression in ageing mouse heart," *Nature*, 2006.

[4] J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, and J. S. Weissman, "Single-cell proteomic analysis of s. cerevisiae reveals the architecture of biological noise," *Nature*, 2006.

[5] S. Di Talia, J. M. Skotheim, J. M. Bean, E. D. Siggia, and F. R. Cross, "The effects of molecular noise and size control on variability in the budding yeast cell cycle," *Nature*, 2007.

[6] K Zhang, J Zhu, J Shendure, GJ Porreca, JD Aach, RD Mitra, and GM Church, "Long-range polony haplotyping of individual human chromosome molecules.," *Nat Genet*, vol. 38, no. 3, pp. 382–7, 2006 Mar.

[7] R.D. Mitra, V.L. Butty, J. Shendure, B.R. Williams, D.E. Housman, and G.M. Church, "Digital genotyping and haplotyping with polymerase colonies.," *Proc Natl Acad Sci U S A*, vol. 100, no. 10, pp. 5926–31, 2003.

[8] M. K. Cheezum, W. F. Walker, and W. H. Guilford, "Quantitative Comparison of Algorithms for Tracking Single Fluorescent Particles," *Biophysical Journal*, vol. 81, no. 4, pp. 2378–2388, 2001.

[9] I.F. Sbalzarini and P. Koumoutsakos, "Feature point tracking and trajectory analysis for video imaging in cell biology," *Journal of Structural Biology*, vol. 151, no. 2, pp. 182–195, 2005.

[10] E. Meijering, I. Smal, and G. Danuser, "Tracking in molecular bioimaging," *Signal Processing Magazine, IEEE*, vol. 23, no. 3, pp. 46–53, 2006.

[11] L. Cortés and Y. Amit, "Efficient annotation of vesicle dynamics in video microscopy," *To appear in IEEE Trans. on PAMI*, 2008.

[12] Y. Amit, D. Geman, and X. Fan, "A coarse-to-fine strategy for multi-class shape detection," *IEEE Trans. on PAMI*, vol. 26, pp. 1606–1621, 2004.

[13] Y. Amit and A. Trouvé, "Pop: Patchwork of parts models for object recognition," *International Journal of Computer Vision, online*, 2007.

[14] Y. Amit, *2D Object Detection and Recognition Models, Algorithms, and Networks*, MIT Press, 2002.

[15] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, "Distance transforms of sampled functions," Tech. Rep., Cornell Computing and Information Science, 2004.

[16] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, New York, NY, USA, 1992.

[17] J. A. Rice, *Mathematical Statistics and Data Analysis*, Duxbury Press, 1994.

[18] R. Deriche, "Using Canny's criteria to derive a recursively implemented optimal edge detector," *The International Journal of Computer Vision*, vol. 1, no. 2, pp. 167–187, May 1987.